Automatic Classification of Sustained Vowels Based on Signal Regularity Measures

Juan M. Miramont[1,2], Gastón Schlotthauer[1,2]

[1]*Laboratorio de Señales y Dinámicas No Lineales, Universidad Nacional de Entre Ríos, Oro Verde, Argentina.*
[2]*Instituto de Investigación y Desarrollo en Bioingeniería y Bioinformática (CONICET-UNER), Oro Verde, Argentina*

***Abstract*— In 1995, Ingo Titze proposed a classification scheme to classify vocalic phonemes in three types (Type I, Type II and Type III), based on the periodicity of the voice signal. Nowadays, voices are classified by the means of spectrograms, although criteria for distinguishing among voice types are not clear yet, especially between Type I and Type II voices. Consequently, there exists great interprofessional variation in the type of voice assigned, and this also depends on each specialist's expertise. As an approach to a more objective classification, features to discriminate between Type I and Type II voices were extracted and then used to classify voices from an annotated dataset. Classic acustic parameters, like Jitter and Shimmer measures, and harmonics to noise ratio (HNR) were used, along with first rahmonic (R1) and two original, here proposed, parameters: principal component normalized variance (VNCP) and spectral peak-valley (PV) ratio. For the classification task, a support vector machine algorithm with linear kernel function was used, feeded with the features that minimized the cross-validation error. An error of 11.61% was obtained, with classification rates of 93.24% and 83.95% for Type I and Type II voices correspondingly.**

***Keywords*— voice types, sustained vowel classification, voice signal processing, support vector machine.**

***Resumen*— En el año 1995 Ingo Titze propuso un sistema de clasificación de fonemas vocálicos en tres tipos (Tipo I, Tipo II y Tipo III) en base a la regularidad de la señal de voz cuasiperiódica correspondiente. En la práctica clínica fonoaudiológica, esta clasificación se realiza en base a la inspección visual de espectrogramas, no siendo claros los criterios que diferencian un tipo de voz de otra, especialmente entre los tipos I y II. En consecuencia, existe una gran variación interprofesional y una fuerte dependencia de la experiencia de cada especialista. Con el fin de lograr una clasificación objetiva basada en parámetros cuantitativos, se buscó extraer características capaces de representar las diferencias fundamentales entre las voces de Tipos I y II, para luego clasificar una base de datos anotada. Se extrajeron parámetros acústicos clásicos, como medidas de Jitter y Shimmer, y harmonics to noise ratio (HNR) calculados utilizando PRAAT. También se propuso la utilización de la amplitud del primer ramónico (R1) y dos características ideadas por los autores de este trabajo: varianza normalizada de la primera componente principal (VNCP) y razones pico-valle (PV) espectrales. La clasificación se realizó mediante máquinas de soporte vectorial (SVM) de kernel lineal utilizando las características que minimizan el error del clasificador. Como resultado, se obtuvo un error de validación cruzada de 11.61%, con porcentajes de acierto del 93.24% y 83.95%, para voces Tipo I y Tipo II respectivamente.**

***Palabras clave*— tipos de voces, clasificación de vocales sostenidas, procesamiento de la señal de voz, máquinas de soporte vectorial.**

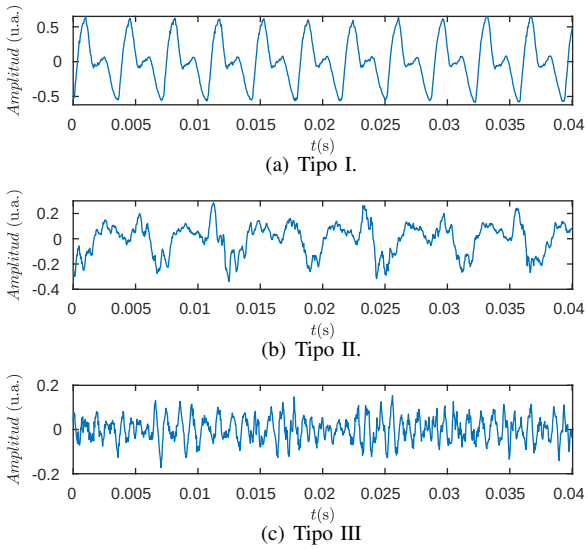
I. Introducción

En el año 1995, Ingo Titze publicó un esquema de clasificación de fonemas vocálicos en tres tipos con el objetivo de definir qué parámetros acústicos de uso clínico (como Jitter, Shimmer y otras medidas de perturbación) pueden determinarse con precisión para cada tipo. De esta manera el especialista en fonoaudiología puede depositar cierta confianza en la estimación de estos parámetros, o no, según el tipo de voz al que pertenezca la vocalización estudiada [1]. A lo largo del presente trabajo, al hablar de señales de voz, se estará haciendo referencia a aquellas correspondientes a la emisión de vocales sostenidas a menos que se indique lo contrario.

Las señales Tipo I poseen una regularidad evidente, sin modificaciones cualitativas en la forma de onda ciclo a ciclo (Fig. 1(a)). Las señales Tipo II se caracterizan por tener frecuencias modulantes o subarmónicas de mayor energía que resultan en fluctuaciones en la forma de onda de la señal a lo largo de los pseudoperiodos (Fig. 1(b)). Finalmente las voces Tipo III no poseen una estructura periódica evidente en el segmento de análisis (Fig. 1(c)).

Contacto: Juan M. Miramont, Gastón Schlotthauer. Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Ruta prov. 11, km 10, Oro Verde, Entre Ríos, Argentina, {jmiramont,gschlotthauer}@conicet.gov.ar

En la actualidad, el uso de espectrogramas de banda angosta es el método más difundido para realizar la clasificación de las voces en la práctica clínica fonoaudiológica, junto con la inspección visual de la representación temporal de la señal y un análisis perceptual. La utilización de espectrogramas tiene la desventaja de depender de varios parámetros involucrados en su representación gráfica, lo que condiciona su interpretación. Si bien los parámetros propuestos por Sprecher y cols. [2] se encuentran bastante difundidos, no existe aún un consenso en



(a) Tipo I.

(b) Tipo II.

(c) Tipo III

Fig. 1: Representación temporal de los distintos tipos de señales de voz correspondientes a fonemas vocálicos según Titze [1].

el área acerca de los parámetros adecuados para la graficación de espectrogramas de señales de voz. Como consecuencia, el uso de esta representación tiempo - frecuencia introduce un factor de subjetividad en la clasificación, pudiendo diferir entre profesionales la clase asignada a una misma señal. Esto resulta en la utilización de herramientas de análisis cuyos resultados podrían no ser confiables, o no representar el estado actual del paciente.

Las voces Tipo I y II pueden ser confundidas si ambas tienen frecuencias modulantes y subarmónicas. La diferencia fundamental entre ambos tipos de voz es la *magnitud* de la energía de estas frecuencias (modulantes y subarmónicas) respecto a la magnitud de la frecuencia fundamental. Como consecuencia, diferenciar entre voces Tipo I y Tipo II es difícil incluso para los especialistas, que a menudo no logran consensuar la clasificación correcta. La separación de las voces Tipo I y II por un lado, de las Tipo III por el otro, no constituye una tarea difícil con las técnicas actuales [4, 20]. Por lo expuesto anteriormente, resulta de interés para el campo de la fonoaudiología clínica el desarrollo de un sistema informático capaz de sugerir al profesional el tipo de voz correspondiente a un paciente, con el fin de volver más objetiva la clasificación de los fonemas vocales. Si bien esta solución se encuentra lejos de concretarse, un primer problema a encarar es la selección de características capaces de describir de manera cuantitativa las diferencias entre los distintos tipos.

En este trabajo se utilizaron características conocidas, junto con otras propuestas por los autores, para clasificar señales en los Tipos I y II empleando un algoritmo de clasificación basado en máquinas de soporte vectorial. Los descriptores extraídos fueron utilizados para evaluar distintos aspectos de la regularidad de la señal, como las modificaciones de amplitud y frecuencia ciclo a ciclo o la conservación de la forma de onda durante el segmento de análisis.

II. Materiales y Métodos

A. Base de datos

Las voces utilizadas en este trabajo constituyen un subconjunto de voces patológicas de la base de datos grabada en el establecimiento *Massachusetts Eye and Ear Infirmary*, en 1994. Esta base de datos ha sido estudiada en varios trabajos [3–5]. Asimismo, el subconjunto empleado ha sido seleccionado y descrito por Parsa y Jamieson [6], y se caracteriza por contar con señales de voz correspondientes a sujetos con distribución etaria similar, y una proporción balanceada de sujetos masculinos y femeninos. Vale aclarar que la clasificación de señales no tiene relación con el sexo del sujeto.

En total se analizaron 155 voces, cuya descripción puede verse en la Tabla I. Cada registro tiene una duración de 1 s, fue adquirido con una frecuencia de muestreo de 44.1 kHz y luego remuestreado a 25 kHz para su almacenamiento. Los registros incluyen únicamente la parte estable de la emisión (las señales ya están preprocesadas). De las voces analizadas, 74 fueron clasificadas como Tipo I y 81 como Tipo II.

Tabla I: Descripción de la población utilizada (voces patológicas) para la base de datos. “M” corresponde a masculino y “F” a femenino.

Cantidad de sujetos		Edad promedio (años)		Rango de edad (años)		Desviación estándar (años)	
M	F	M	F	M	F	M	F
68	87	41.8	37.4	26-58	21-51	9.3	8.1

Las voces fueron clasificadas por una profesional de la fonoaudiología en los tres tipos propuestos por Titze, y luego se utilizó este corpus anotado como referencia para medir el desempeño del algoritmo clasificador.

B. Características utilizadas

1) Jitter y Shimmer: El Jitter es un fenómeno definido como una perturbación de corto plazo (ciclo a ciclo) de la frecuencia fundamental de la voz o, de forma equivalente, del periodo fundamental. El fenómeno del Shimmer, de manera análoga, se describe como una perturbación de corto plazo pero de la amplitud pico a pico de la señal [7].

Para medir estos fenómenos, se emplearon la razón de Jitter porcentual ($Jitter_{\%}$) y razón de Shimmer porcentual ($Shimmer_{\%}$) definidas como:

$$Jitter_{\%} = 100 \frac{\frac{1}{N-1}\sum_{n=2}^{N} |T_n - T_{n-1}|}{\frac{1}{N}\sum_{n=1}^{N} T_n}, \quad (1)$$

$$Shimmer_{\%} = 100 \frac{\frac{1}{N-1}\sum_{n=2}^{N} |A_n - A_{n-1}|}{\frac{1}{N}\sum_{n=1}^{N} A_n}, \quad (2)$$

donde $T_n = \{T_1, T_2, ..., T_N\}$ y $A_n = \{A_1, A_2, ..., A_N\}$ son la serie de periodos y la serie de amplitudes de la señal, respectivamente, siendo N el número de elementos de cada una.

2) Harmonics to Noise Ratio (HNR): La razón entre la energía de la componente armónica y la energía de la componente ruidosa, *Harmonics to Noise Ratio* (HNR), es una medida propuesta por Yumoto y cols. [8] que busca cuantificar el grado de periodicidad de una señal $x(t)$. Para el cálculo de HNR se utilizó el software PRAAT de Paul Boersma y David Weenink [9]. El método publicado por Boersma en 1993 [10] estima el valor de HNR mediante:

$$HNR = 10 \log \left(\frac{r_x(T_0)}{r_x(0) - r_x(T_0)} \right). \quad (3)$$

donde T_0 es el periodo fundamental y $r_x(t)$ es la autocorrelación de la señal $x(t)$.

3) Primer Ramónico (R1): El primer *ramónico* es el primer pico prominente en el cepstrum real [11, 12] de una señal cuasiperiódica. Para las señales de voz correspondientes a fonemas vocálicos, este pico se localiza en el periodo fundamental.

El cepstrum real de una señal discreta $x[n]$ se define como:

$$c[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \left| X(e^{j\omega}) \right| e^{j\omega n} d\omega, \quad (4)$$

donde $X(e^{j\omega})$ es la transformada de Fourier de tiempo discreto (DTFT, por sus siglas en inglés) de $x[n]$. Otros autores [13, 14] reportaron que la amplitud de este pico es directamente proporcional a la media geométrica del HNR. Sin embargo, lo atractivo de este indicador es la posibilidad de calcularlo sin necesidad de una estimación de la frecuencia fundamental.

4) Varianza Normalizada de la Componente Principal: Esta característica fue propuesta por los autores de este trabajo y consiste en calcular el porcentaje de varianza explicado por la primera componente principal del conjunto de todos los ciclos de la señal. Para calcular este índice, en primer lugar se segmentó la señal en cada pseudoperiodo, denotados como $\boldsymbol{b}_i$, siendo N_i el número de muestras de cada uno.

El paso siguiente consistió en "estirar" cada segmento $\boldsymbol{b}_i$ a la longitud del segmento de mayor duración, de manera tal que todos tengan el mismo número de muestras. Con este objetivo, se buscó el periodo de mayor duración, cuyo número de muestras es N_{max}. Luego se aumentó el número de muestras de aquellos $\boldsymbol{b}_i$ para los que $N_i \leq N_{max}$ mediante interpolación por splines.

Utilizando estos segmentos como columnas, se conformó una matriz a la que se le realizó un análisis de componentes principales (PCA) [15]. La primera componente principal hallada, que podría considerarse como la forma de onda más representativa de la señal, es responsable de un mayor porcentaje de la varianza cuanto mayor sea la regularidad de la señal, es decir, cuanto más parecidos sean los ciclos entre sí.

El valor reportado finalmente es el porcentaje de la varianza total debido a la primera componente principal, calculado como:

$$VNCP_{\%} = \frac{\lambda_1}{\sum_{i=1}^{M} \lambda_i} \times 100\%, \quad (5)$$

donde λ_1 es la varianza de la primera componente principal, λ_i es la varianza de la i-ésima componente principal, y M el número de componentes calculadas.

El objetivo de esta característica es evaluar la conservación de la morfología de la señal. Al comparar los segmentos interpolados, se pierde la información acerca de la variación del periodo de la señal pero es posible comparar la forma de onda en cada ciclo. Cuanto mayor sea el valor de VNCP, mayor será la similitud entre la morfología de la señal en cada ciclo.

5) Razón Pico-Valle (PV): El espectro de una señal de voz correspondiente a la emisión de una vocal sostenida se caracteriza por poseer *picos*, ubicados en múltiplos enteros de la frecuencia fundamental, y *valles* entre dichos picos. Se espera que, para señales poco regulares y con mayor nivel de ruido, los picos del espectro sean más anchos y los valles menos profundos. Esto podría deberse a la presencia de frecuencias subarmónicas de gran magnitud en la señal, como ocurre particularmente para las voces Tipo II, que aumenten el nivel de energía en los valles y disminuyan así la prominencia de los picos. Por el contrario, para señales más regulares y con frecuencias subarmónicas de menor magnitud, se espera que la energía del espectro se encuentre más concentrada en las frecuencias armónicas, y los valles sean más profundos.

Bajo esta hipótesis, la razón entre la energía en los picos del espectro y la energía de los valles inmediatamente a su izquierda, debería ser mayor para señales Tipo I que para las señales Tipo II.

El i-ésimo cociente Pico-Valle, correspondiente a la razón entre la energía del i-ésimo pico y i-ésimo valle, se define como:

$$PVi = \frac{\sum_{r \in P_i} |X[r]|^2 (N_{P_i})^{-1}}{\sum_{r \in D_i} |X[r]|^2 (N_{D_i})^{-1}}$$

donde $X[r]$ es la transformada de Fourier discreta (DFT, por sus siglas en inglés) de la señal analizada $x[n]$, N_{P_i} y N_{D_i} son el número de puntos en la región P_i y D_i, pico y valle respectivamente.

C. Máquinas de Soporte Vectorial

Las máquinas de soporte vectorial (SVM, por sus siglas en inglés) son un conjunto de algoritmos desarrollados en base a las ideas aportadas por Vapnik y Cortes [16, 17]. Su objetivo consiste en separar dos clases dejando el mayor margen posible entre ellas. El algoritmo transforma el espacio de las características de entrada en otro espacio (que puede ser de mayor dimensión) en el que separa las clases mediante un hiperplano. Este hiperplano es determinado por un subconjunto de los datos, llamados *vectores soporte*. La función discriminante de una máquina de soporte vectorial es:

$$g(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) + w_0, \quad (6)$$

donde $\boldsymbol{x}$ es el dato a clasificar, N es el número de datos y K es una función escalar conocida como función *kernel* o núcleo. Los valores de w_0 y α_i deben ser determinados para cada problema. La función núcleo utilizada en este trabajo se conoce como kernel lineal:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j. \quad (7)$$

Se descartaron otras funciones núcleo, como el kernel gaussiano, debido a que no se obtuvieron mejores resultados que con el kernel lineal.

D. Area bajo la curva ROC (AUC)

Las gráficas ROC (del inglés *Receiver operating characteristics*) [18] son gráficos de *sensibilidad* vs. 1-*especificidad* para un sistema clasificador *binario*. Las curvas ROC son útiles para evaluar clasificadores, visualizar su desempeño y seleccionar el más adecuado de un conjunto.

Para la selección de un clasificador, suele utilizarse como parámetro el área bajo la curva ROC (AUC). Este valor se encuentra entre 0 y 1. Cuanto más cerca de la unidad se encuentre, mejor será el clasificador evaluado. Esto se debe a que un clasificador con un área cercana a 1 posee puntos de operación cercanos al $(0, 1)$, que se corresponde con el clasificador perfecto.

Asimismo, el área bajo la curva correspondiente a una característica dada es la probabilidad de que un dato de una clase tomado aleatoriamente posea, para esa característica, un valor mayor que para un dato de la otra clase, también elegido de forma aleatoria.

Como regla práctica, debe elegirse el clasificador que posea el mayor valor de área bajo la curva ROC. Si hubiera áreas menores a 0.5, se debe invertir la salida del clasificador y luego evaluar el AUC nuevamente.

E. Selección Secuencial de Características

La *selección* de características se utiliza para reducir el número de descriptores utilizados en un problema, y se basa en seleccionar aquellas características que maximicen el desempeño del clasificador. El método consiste, en primer lugar, en ordenar las características por su poder de discriminación individual, utilizando, por ejemplo, el área bajo la curva ROC. De esta manera se pueden descartar aquellas con menor capacidad de separación entre clases (las últimas listadas).

En segundo lugar, se deben eliminar aquellos descriptores que aportan información redundante. Para esto es posible calcular los índices de correlación entre todas las características ordenadas o un subconjunto menor de éstas, descartando aquellas con un índice de correlación alto.

Finalmente, se debe evaluar el desempeño del clasificador con diferentes subconjuntos de las características no descartadas para conservar aquellas responsables del mayor desempeño. Una manera de hacerlo es realizar una selección en forma secuencial, agregando (*búsqueda hacia delante*) o removiendo (*búsqueda hacia atrás*) características hasta alcanzar alguna condición de parada, que generalmente suele ser un umbral de desempeño del clasificador.

III. Resultados

A. Gráficos de cajas de las características

Las características utilizadas fueron extraídas con el objetivo de diferenciar las voces Tipo I y Tipo II. En otras palabras, estos indicadores deben tomar valores distintos para los distintos tipos de voz analizados. De los gráficos de cajas de las Figs. 2, 3, 4, 5, 6 y 7 se observa que las características difieren en el rango de valores tomado para los dos tipos de voz analizados en este trabajo.

Con respecto a los cocientes pico-valle del espectro, sólo los valores obtenidos para el primer cociente (PV1) mostraron una diferencia estadísticamente significativa entre los distintos tipos de voz. Para esta determinación se utilizó el test no paramétrico de suma de rangos con signo de Wilcoxon, obteniéndose un valor $p < 0.001$.

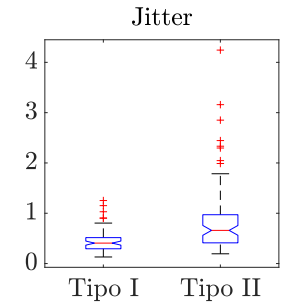


Fig. 2: Gráfico de cajas que muestra la distribución de los valores de la característica Jitter$_{\%}$, para voces Tipo I y Tipo II..

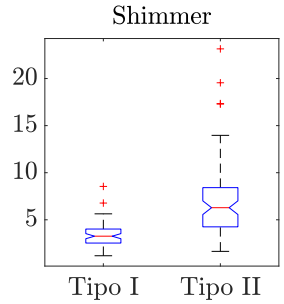


Fig. 3: Gráfico de cajas que muestra la distribución de los valores de la característica Shimmer$_{\%}$, para voces Tipo I y Tipo II..

B. Área bajo la curva ROC

La Tabla II reporta el área bajo la curva ROC de las características extraídas, ordenadas de mayor a menor. La característica con mayor área bajo la curva es VNCP, con 0.9394. Por lo tanto, es la característica con mayor poder de discriminación, por sí sola, entre las voces Tipo I y Tipo II. Asimismo, es posible observar que los mayores valores de AUC se corresponden con las características que mayor separación tienen entre los rangos de valores que toman para las voces Tipo I y las voces Tipo II.

Tabla II: Características ordenadas por mayor área bajo la curva ROC.

1	VNCP	0.9394
2	HNR	0.8931
3	Shimmer$_{\%}$	0.8410
4	R1	0.7819
5	Jitter$_{\%}$	0.7436
6	PV1	0.6708

C. Correlación entre las características

La Tabla III muestra la matriz de coeficientes de correlación lineal entre las variables extraídas. Se resaltan en gris los coeficientes más grandes. No obstante, no se descartó ninguna variable con este análisis debido a que los coeficientes no permiten establecer con claridad la existencia de una correlación importante.

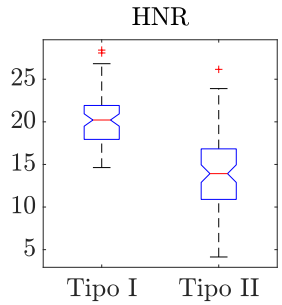


Fig. 4: Gráfico de cajas que muestra la distribución de los valores de la característica HNR para voces Tipo I y Tipo II..

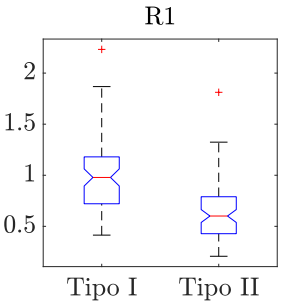


Fig. 5: Gráfico de cajas que muestra la distribución de los valores de la característica R1 para voces Tipo I y Tipo II..

Tabla III: Coeficientes de correlación entre las características extraídas.

	R1	HNR	Jitter	VNCP	PV1	$\text{Shimmer}_{\%}$
R1	1.000	0.727	-0.474	0.440	0.084	-0.545
HNR	0.727	1.000	-0.596	0.702	0.082	-0.747
Jitter	-0.474	-0.596	1.000	-0.515	-0.162	0.754
VNCP	0.440	0.702	-0.515	1.000	0.241	-0.576
PV1	0.084	0.082	-0.162	0.241	1.000	-0.201
$\text{Shimmer}_{\%}$	-0.545	-0.747	0.754	-0.576	-0.201	1.000

D. Selección secuencial de características

Teniendo en cuenta el orden de características de la Tabla II, se agregaron una a una al vector de características que alimenta a la máquina de soporte vectorial. Cada vez que se agrega una característica se calcula el error de validación cruzada de 10 iteraciones. Si el error calculado disminuye se conserva la característica. Si no, se descarta y se agrega la que sigue en la Tabla II. Se reportan en la Tabla IV los errores calculados.

Tabla IV: Desempeño de máquina de soporte vectorial con núcleo lineal para cada conjunto de características.

Error	VNCP	VNCP+PV1	$\text{VNCP+PV1+Shimmer}_{\%}$
Leave-One-Out	14.84%	13.55%	11.61%
V. C. de 10 it.	15.00%	13.59%	11.61%

El menor error obtenido fue de 11.61%, con la combinación de características VNCP, PV1 y $\text{Shimmer}_{\%}$. Cualquier combinación de cuatro características o más no redujo el error a la salida del clasificador, por lo tanto fueron descartadas. También se realizó una búsqueda exhaustiva de la mejor combinación de características, que consistió en calcular el desempeño del clasificador para todas las combinaciones posibles de características y seleccionar la de mayor desempeño.

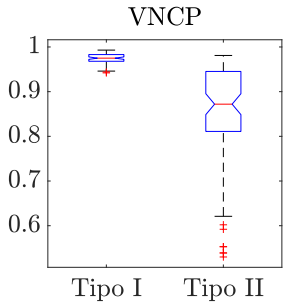


Fig. 6: Gráfico de cajas que muestra la distribución de los valores de la característica VNCP para voces Tipo I y Tipo II..

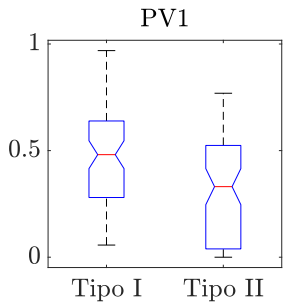


Fig. 7: Gráfico de cajas que muestra la distribución de los valores de la característica PV1 para voces Tipo I y Tipo II.

La combinación seleccionada con este último método fue VNCP, PV1 y $\text{Shimmer}_{\%}$, corroborando el resultado obtenido por selección secuencial.

Para la mejor combinación de características, se muestra en la Tabla V la matriz de confusión del clasificador junto con la sensibilidad, la especificidad y el error de validación cruzada de 10 iteraciones ya reportado. La sensibilidad, en este caso, se corresponde con el porcentaje de aciertos del clasificador para las voces Tipo I. Asimismo, la especificidad se corresponde con el porcentaje de aciertos para las voces Tipo II.

Tabla V: Matriz de confusión del clasificador SVM para la combinación *VNCP+PV1+*$\textit{Shimmer}_{\%}$.

	Salida del Clasificador	
	Tipo I	Tipo II
Tipo I	69	5
Tipo II	13	68
$\text{Error}_{V.C.10it.}$	11.61%	
Sensibilidad	93.24%	
Especificidad	83.95%	

E. Comparación con Trabajos Previos

En 2016 Ji Yeoun Lee [19] propuso utilizar estadísticos de orden superior junto con parámetros acústicos, como medidas de Jitter, Shimmer y relación señal - ruido (SNR), para la clasificación de voces correspondientes a fonemas vocálicos en los tipos descriptos. Los estadísticos de orden superior utilizados en [19] fueron: coeficiente de variación de asimetría normalizada (CSV, por sus siglas en inglés), coeficiente de variación de kurtosis normalizada (CKV, por sus siglas en inglés) y valor de bicoherencia (BV, por sus siglas en inglés).

Los resultados reportados en [19] se resumen en la Tabla VI. La base de datos utilizada en [19] difiere de la empleada en el presente trabajo (cuenta con 70 voces, 35 de cada tipo, de sujetos de origen surcoreano) por lo que una comparación directa no es posible.

Con el objetivo de comparar los resultados sobre un mismo conjunto de datos, se extrajeron las características utilizadas en [19] (Shimmer$_{\%}$, Jitter$_{\%}$, SNR, CSV, CKV y BV) de las 155 voces empleadas en el presente trabajo. La Tabla VII muestra los porcentajes de sensibilidad y especificidad obtenidos al utilizar dichas características y una SVM de kernel lineal para la clasificación (se utilizó HNR en lugar de SNR).

Al comparar la sensibilidad y la especificidad reportadas en las Tablas V y VII, es posible observar que las medidas propuestas por los autores de este trabajo (VNCP y PV1) en combinación con Shimmer$_{\%}$, parecen describir mejor las diferencias entre las voces Tipo I y Tipo II que las propuestas en [19], obteniéndose un mejor desempeño del algoritmo clasificador.

Tabla VI: Resultados obtenidos por Ji Yeoun Lee en [19].

	Jitter$_{\%}$+Shimmer$_{\%}$+SNR	Jitter$_{\%}$+Shimmer$_{\%}$+SNR +CSV+CKV+BV
Sensibilidad	62.86%	85.75%
Especificidad	74.28%	85.75%

Tabla VII: Resultados obtenidos al extraer las características propuestas en [19] del conjunto de 155 voces empleado en el presente trabajo, utilizando una SVM de kernel lineal para la clasificación.

	Jitter$_{\%}$+Shimmer$_{\%}$+HNR +CSV+CKV+BV
Sensibilidad	87.84%
Especificidad	76.54%

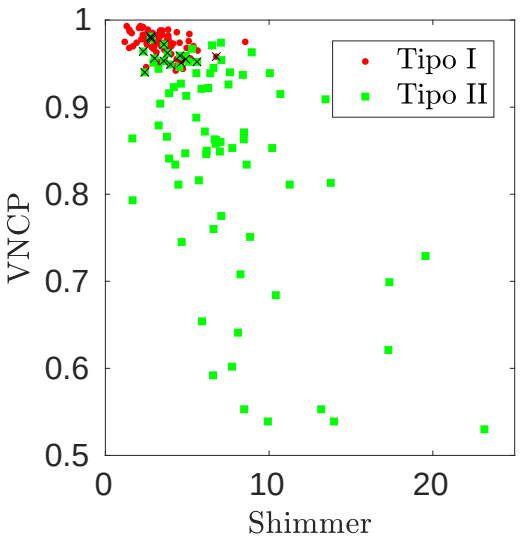


Fig. 8: Gráfico de dispersión de los datos para las características VNCP y Shimmer. Los datos con una "x" fueron clasificados incorrectamente.

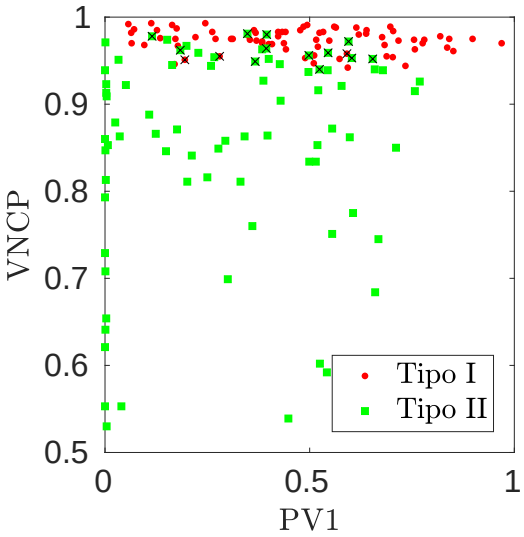


Fig. 9: Gráfico de dispersión de los datos para las características VNCP y Shimmer. Los datos con una "x" fueron clasificados incorrectamente.

IV. Discusión

Con el objetivo de clasificar automáticamente señales de voz correspondientes a fonemas vocálicos se implementó un algoritmo basado en la extracción de características y la utilización de máquinas de soporte vectorial de kernel lineal. De las seis características extraídas, sólo se utilizaron tres para obtener la mejor clasificación: VNCP, PV1 y Shimmer$_{\%}$, seleccionadas por minimizar el error del clasificador.

La característica VNCP, propuesta por los autores de este trabajo, fue la de mayor capacidad de discriminación individual como reflejan los valores de AUC. Esta característica busca evaluar el grado de la regularidad de la señal, segmentándola en sus periodos y aplicando análisis de componentes principales para encontrar una forma de onda representativa de la señal. Los valores obtenidos para este indicador sostienen esta afirmación, tendiendo a ser mayores para las señales Tipo I, más regulares y con pequeñas perturbaciones ciclo a ciclo; que para las señales Tipo II, con fluctuaciones de mayor magnitud entre cada periodo.

Es importante destacar que, a través de VNCP, se buscó evaluar si la forma de onda de cada ciclo es similar a lo largo de todo el segmento de análisis como una medida de la regularidad de la señal. Otras características, como Jitter$_{\%}$ o Shimmer$_{\%}$, buscan cuantificar la regularidad de la señal desde otras perspectivas, como la variación de la frecuencia fundamental o la variación de la amplitud máxima ciclo a ciclo.

El valor de especificidad del 83.95% obtenido (Tabla V) indica que aún existen dificultades para diferenciar las voces Tipo II cuando son muy parecidas a las Tipo I. Esto también se evidencia en los gráficos de dispersión de las Figs. 8 y 9. Además, de las mismas figuras se observa que el conjunto de las voces Tipo II es más disperso, lo que implica que los valores que toman las características para este tipo de voces son mucho más variables que los que toman las voces Tipo I. Esto coincide con lo observado por otros autores [4, 19, 20].

Se observó que el desempeño del algoritmo clasifi-

cador fue mejor al emplear los descriptores aquí propuestos (VNCP, Shimmer$_{\%}$, PV1), en comparación con los resultados obtenidos al utilizar las características sugeridas en [19] (BV, CKV, CSV, Jitter$_{\%}$, Shimmer$_{\%}$, HNR), extraídas del mismo conjunto de voces.

V. Conclusiones

Se realizó un algoritmo para la clasificación de señales de voz correspondientes a fonemas vocálicos extrayendo tres características, VNCP, PV1 y Shimmer$_{\%}$, utilizando una máquina de soporte vectorial de kernel lineal. Se obtuvo un error de validación cruzada de 11.61%, con porcentajes de acierto de 93.24% y 83.95%, para voces Tipo I y Tipo II respectivamente.

Estos resultados indican que el conjunto de características empleado representa las diferencias entre las voces Tipo I y Tipo II, aunque con limitaciones. Las características utilizadas se presentan como una primera aproximación para traducir los parámetros subjetivos utilizados actualmente, en parámetros objetivos y cuantificables.

Por otro lado, si bien los resultados obtenidos se corresponden en gran medida con la clasificación llevada a cabo por una profesional de la fonoaudiología, debe tenerse en cuenta que aún no existe un consenso en la práctica clínica de esta especialidad que permita realizar una clasificación objetiva. Como consecuencia, el tipo de voz asignado a una misma señal podría variar entre profesionales. Por esa razón, la participación de varios especialistas es clave para validar las características extraídas y el algoritmo clasificador en su conjunto.

Se destaca el desempeño obtenido mediante la característica VNCP que fue propuesta originalmente para este trabajo, y fue la característica con mejor capacidad de discriminación individual de todas las estudiadas. Su utilización como un método para obtener una forma de onda representativa de un fenómeno aproximadamente periódico, y su capacidad de cuantificar el grado de regularidad de la señal analizada, podrían ser útiles en otros problemas como la clasificación de electroglotogramas.

En trabajos posteriores se buscará comparar el desempeño del algoritmo presentado en este trabajo con los propuestos por otros autores, aplicados sobre el mismo conjunto de datos.

Agradecimientos

Este trabajo fue realizado gracias al apoyo de la Universidad Nacional de Entre Ríos (UNER), a través del PID 6171, y del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) a través de los proyectos PICT 2012-2954 y PIO 146-201401-00014-CO (CONICET-UNER). Los autores desean extender su agradecimiento a la Lic. Juliana Codino por su colaboración en la clasificación de las señales utilizadas en este trabajo.

Referencias

[1] I. R. Titze, *Workshop on acoustic voice analysis: Summary statement*. National Center for Voice and Speech, 1995.

[2] A. Sprecher, "Updating signal typing in voice: addition of type 4 signals," *The Journal of the Acoustical Society of America*, vol. 127, no. 6, pp. 3710–3716, 2010.

[3] J. D. Arias-Londoño, J. I. Godino-Llorente, M. Markaki, y Y. Stylianou, "On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices," *Logopedics Phoniatrics Vocology*, vol. 36, no. 2, pp. 60–69, 2011.

[4] W. M. Calawerts, L. Lin, J. Sprott, y J. J. Jiang, "Using rate of divergence as an objective measure to differentiate between voice signal types based on the amount of disorder in the signal," *Journal of Voice*, 2016.

[5] P. Henríquez, J. B. Alonso, M. A. Ferrer, C. M. Travieso, J. I. Godino-Llorente, y F. Díaz-de María, "Characterization of healthy and pathological voice through measures based on nonlinear dynamics," *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 6, pp. 1186–1195, 2009.

[6] V. Parsa y D. G. Jamieson, "Identification of pathological voices using glottal noise measures," *Journal of Speech, Language, and Hearing Research*, vol. 43, no. 2, pp. 469–485, 2000.

[7] R. J. Baken y R. F. Orlikoff, *Clinical measurement of speech and voice*. Cengage Learning, 2000.

[8] E. Yumoto, W. J. Gould, y T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *The journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1544–1550, 1982.

[9] P. Boersma y V. van Heuven, "Speak and unspeak with PRAAT," *Glot International*, vol. 5, no. 9-10, pp. 341–347, 2001.

[10] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*, vol. 17, 1993.

[11] B. P. Bogert, M. J. Healy, y J. W. Tukey, "The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Proceedings of the symposium on time series analysis*, vol. 15. chapter, pp. 209–243, 1963.

[12] A. V. Oppenheim y R. Schafer, "Digital signal processing," *Prentice-Hall, Englewood Cliffs, New Jersey*, vol. 6, pp. 125–136, 1975.

[13] A. Alpan, J. Schoentgen, Y. Maryn, F. Grenez, y P. Murphy, "Assessment of disordered voice via the first rahmonic," *Speech Communication*, no. 5, pp. 655–663, 2012.

[14] P. J. Murphy, "On first rahmonic amplitude in the analysis of synthesized aperiodic voice signals." *The Journal of the Acoustical Society of America*, vol. 120, no. 5 Pt 1, pp. 2896–2907, 2006.

[15] C. M. Bishop, "Pattern recognition," *Machine Learning*, vol. 128, 2006.

[16] V. Vapnik, *The Nature of Statistical Learning Theory*, ser. Information Science and Statistics. Springer New York, 1999.

[17] C. Cortes y V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[18] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[19] J. Lee, "Parameter estimations for signal type classification of korean disordered voices," *International Journal of Engineering and Technology*, vol. 7, no. 6,

pp. 1977 – 1988, 2016.

[20] L. Lin, W. Calawerts, K. Dodd, y J. J. Jiang, "An objective parameter for quantifying the turbulent noise portion of voice signals," *Journal of Voice*, 2015.

Juan Manuel Miramont nació en San Miguel, Provincia de Buenos Aires, en 1992. Se recibió de Bioingeniero de la Universidad Nacional de Entre Ríos (UNER) y comenzó su doctorado en ingeniería en la Universidad Nacional del Litoral en el año 2017. Desde el año 2015 se desempeña como becario en el Laboratorio de Señales y Dinámicas no Lineales de la Facultad de Ingeniería (UNER).

Gastón Schlotthauer nació en General Galarza, Entre Ríos, en 1974. Recibió los títulos de Bioingeniero y Magíster en Ingeniería Biomédica de la Universidad Nacional de Entre Ríos (UNER), Argentina, en 2000 y 2007, respectivamente, y el de Doctor en Ingeniería de la Universidad Nacional del Litoral, Argentina, en 2010. Se unió al Laboratorio de Señales y Dinámicas no Lineales y al Departamento de Matemática (Facultad de Ingeniería, UNER) en 1995 y 1998 respectivamente. Desde 2011 es investigador de CONICET, y desde 2017 es Profesor Titular Ordinario en la UNER. Sus intereses en investigación incluyen procesamiento de señales biomédicas, procesamiento adaptativo de señales, dinámicas no lineales y aprendizaje maquinal, entre otros tópicos.